ARMY RESEARCH LABORATORY

Probabilistic Modeling of a Syndrome

by Josip Z. Šoln

ARL-TR-1268

December 1996

Approved for public release; distribution is unlimited.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer need. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground (EA), MD 21010-5423

ARL-TR-1268 **December 1996**

Probabilistic Modeling of a Syndrome

Josip Z. Šoln
Survivability/Lethality Analysis Directorate, ARL

Abstract

We propose a probabilistic methodology for deducing a syndrome or syndromes (possibly induced by chemical/biological agents) associated with a large number of people from certain geographic areas that have well-established diagnoses and symptoms. Here, using the finite element method and the databases of symptoms and diagnoses, for each geographic area an analytical probability distribution function is established, which gives a probability that a person has a certain number of symptoms/diagnoses. This, in turn, allows us to write down an analytic expression for the symptoms/diagnoses density from which, with the help of databases, one deduces the overall most numerous symptoms and diagnoses; as such, they define the syndrome for the particular geographic area. Now, comparing the syndromes to each other, one can see to what extent geography, and what is on it, affects the syndromes associated with different geographic areas.

ACKNOWLEDGMENTS

The author wishes to thank Al Gluckman for explaining his report [1] and for very many useful discussions. Thanks are due to Ron Weiss and Lynn Davis for their patient and kind explanations of medical and chemical terminologies, respectively. The interest and many valuable suggestions by Brian Ruth are gratefully acknowledged.

INTENTIONALLY LEFT BLANK.

TABLE OF CONTENTS

INTENTIONALLY LEFT BLANK.

LIST OF FIGURES

INTENTIONALLY LEFT BLANK.

1. INTRODUCTION

It happens quite often that the problem, whose resolution one pursues inductively, will yield a large number of possible solutions, which are only vaguely related. Then, of course, one is faced with a new problem: Which of these solutions represents the one that is realistic for the situation at hand? A possible way to circumvent this dilemma is to start deductively rather than inductively from the very beginning, which, by its very nature, should lead to just one solution of the problem at hand. Such a situation, we believe, is found when one is trying to deduce a syndrome from a collection of symptoms and diagnoses for a large number of people that were confined to a specific geographic area. Here, we are inspired by the so-called Gulf War and reported cases of symptoms of the U.S. military personnel that were acquired during the Iraq-Kuwait War of 1991. The idea here is to have a deductive methodology to deal with this type of situation in the future.

We believe that the probabilistic methodology we will describe should be able to yield syndromes that one could uniformly associate with particular geographies or particular medical preventive procedures (e.g., inoculation). As such, it represents a deductive approach for finding out the syndrome or syndromes, if there are more than one.

What we intend to do here is to give the mathematical aspects of the probability distribution function of symptoms/diagnoses for a soldier associated with some military unit in a particular geographic location. The importance of this distribution function lies in the fact that all the soldiers are now treated "equally" in a sense that the distribution function tells us what the probability is for any soldier to have, say, five symptoms/diagnoses of any kind. Of course, the data, which are compiled and will be available from the U.S. Army and Joint Services Environmental Support Group and the Veterans Administration Epidemiology Services, will make these distribution functions relevant for particular units and particular geographic locations.

Needless to say, the methodology that employs the probabilistic formulation should also be applicable to other situations where a large number of people from a certain geographic area have well-established illnesses and symptoms. As such, it may go beyond the military applications and, hopefully, establish itself as a useful tool in preventing the spread of illnesses due to abuses of the environment.

In the first part of section 2, we give a simplified description of necessary data, i.e., symptoms, diagnoses, syndromes, unit identification code (UIC), Julian date (J-date), and the military grid reference system (MGRS). The second part of section 2 is devoted to casting these data into a simplified form suitable for using them in formulations of probability distribution functions. Section 3 is devoted to the probabilistic formulation of the syndrome with the help of finite element method. Discussion, recommendations, and conclusion are in section 4.

2. THE DATA

No matter which way we go to deduce a syndrome, one needs the data that, at the end, will make any formulation more or less useful. In our case, the data, of course, should contain both medical and personnel records. Here we are giving a brief description of these data, which, when acquired, should be easy to incorporate into the described methodology. Examples of syndrome descriptions that we finally present, of course, are worked out with fake data merely for the sake of illustrating the probabilistic formulation of syndromes.

2.1 Data Terminology. Here we shall familiarize ourselves with the terminology of data that come from two main sources: The Veterans Administration Epidemiology Services (VA-ES) and the U.S. Army and Joint Services Environmental Support Group (AJS-ESG). For the sake of simplicity, from now on these two services will be referred to simply as VA-ES and AJS-ESG, respectively. We use capital letters to describe these data.

From VA-ES, the data necessary to describe the probabilistic formulation of the syndrome are:

IDENTIFICATION NUMBER - A four-digit number identifying a person.

BRANCH OF SERVICE - Army, Navy, Marines, and Air Force.

THREE INDEPENDENT SYMPTOMS - Each symptom is characterized by a unique string of digits recorded as a code number. Before being specified, these can be denoted as ICDSYM 1, ICDSYM 2, and ICDSYM 3.

THREE INDEPENDENT DIAGNOSES - Each diagnosis is characterized again by a unique string of digits recorded as a code number. And again, before being specified, they can be denoted as ICDIAG 1, ICDIAG 2, and ICDIAG 3.

Both symptoms and diagnoses use the standard ICD-9-CM codes [1]. In the VA codes, the decimal points have been entirely eliminated, so the coded data appear from 001 (for cholera) to the number 9999; hence, 780.7 appears as 7807, for example. When ICDSYM 3 = 0 for a particular person, that person has only two symptoms. Similarly, if for some person ICDIAG 2 = 0 and ICDIAG 3 = 0, that person has only one diagnosis. As an example, a complete symptom and diagnosis description for a person could look something like this:

$$\begin{gathered}\text{ICDSYM 1} = 78999,\ \text{ICDSYM 2} = 7807,\ \text{ICDSYM 3} = 0;\\ \text{ICDIAG 1} = 71940,\ \text{ICDIAG 2} = 0,\ \text{ICDIAG 3} = 0.\end{gathered} \tag{1}$$

When doing mathematical manipulations, these notations would be too cumbersome; so we further simplify these notations as:

$$s_i = \text{ICDSYM i},\ d_j = \text{ICDIAG j},\ i, j = 1, 2, 3. \tag{2}$$

Now, symptoms and diagnoses are used together when describing the illness; therefore, to simplify matters, from now on, we will call symptoms and diagnoses simply generalized symptoms. Furthermore, we may as well put them together in a six-component generalized symptom/diagnosis vector:

$$\phi = (\phi_i) = (s_i, s_2, s_3; d_1, d_2, d_3). \tag{3}$$

The example from the relation (1) now simply reads as $\phi = (78999, 7807, 0; 71940, 0, 0)$, which, of course, gives a complete description of someone's illness. Now, the question is how to arrange these generalized symptoms into a six-component vector. As far as our statistical model is concerned, this arrangement is arbitrary. However, a particular person putting together the probabilistic formulation of a syndrome could arrange them according to their severities (i.e., s_1 is the most severe symptom, while s_2 is less, and s_3 the least severe symptom). Similarly, we do the same with the diagnoses—d_1, d_2, and d_3 are in the descending order of their severities. Of course, the person himself would decide how to define the severity degrees of symptoms and diagnoses.

As we see then, the maximum number of generalized symptoms that a person can be assigned is six. In this connection, we introduce a continuous generalized symptom number variable x; its values are restricted between 0 and 6 for a person. Of course, its values may exceed 6 if we are talking about a syndrome that may in its definition have more than six generalized symptoms. In any case, $x = 1$ means that a person has just one generalized symptom; this symptom may be simply $s_1 = 7807$ or $d_2 = 71840$ but not 0, since 0 means no symptoms.

UIC is a unit identification code. Unfortunately, this code is different for different branches of the Service. For the Army, the UIC is a string of numbers preceded by a letter. In general, we shall simplify discussion by skipping the codes and referring to military units simply as UIC1, UIC2, etc.

DATES - Refer to when a person (with an ID number) entered and exited the War Theater associated with a particular geography or geographies. Both of these dates are given in MM/DD/YY.

The VA-ES data are complemented with the data from AJS-ESG. The ones that are needed for the probabilistic treatment of the syndrome are again listed in capital letters:

UIC - It is the same as described by the Veterans Administration database (VA-ES).

JULIAN DATES (J-DATE) - These dates describe when some UICs entered and exited a particular geographic area. The Julian date is a character string YYDDD. If YY = 91, then this corresponds to 1991, and DDD is a particular day in 1991. DDD runs from 001 to 365.

Military Grid Reference System (MGRS) [1] - This is simply a military two-dimensional coordinate frame given as grids. It is designated with two letters and two numbers. The number is always associated with the letter; so if the letter is L, then the number symbolically can be written as N_L. Hence the grid area can be written symbolically as LTN_LN_T. The following situations should be distinguished: (1) $N_{L,T}$ are absent, the grid is 100 km by 100 km; (2) $N_{L,T} = 0, 1, ..., 9$, the grid is 10 km by 10 km; (3) $N_{L,T} = 0, .., 99$, the grid is 1 km by 1 km; (4) $N_{L,T} = 0, ..., 999$, the grid is 100 m by 100 m; and (5) $N_{L,T} = 0, ..., 9999$, the grid is 10 m by 10 m. The first set of digits, N_L in our example, is called easting, while the second set of digits, N_T in our example, is called northing; the grid is determined from its southwest corner.

At the National Institute of Health (NIH) Conference in April 1994, the sets of generalized symptoms (symptoms plus diagnoses) were established for a variety of syndromes. The following generalized symptoms are the ones that might be the most relevant for the probabilistic syndrome study:

(1) Effects of depleted uranium

(2) Effects of pesticides

(3) Effects from petrochemicals and petrochemical smokes

(4) Effects of nerve and mustard gas

(5) Effects of Leishmaniasis infection, and

(6) Effects of other chemicals.

2.2 Assigning the Data. To be able to have a realistic probabilistic formulation of the syndrome, we have to be very careful how the data is assigned. For one thing, we must have a sufficient number of soldiers with the symptoms. Another worry is the time factor. If the symptoms and diagnoses are associated with a particular geographic area, we have to be sure that the unit (or units comprising a larger unit) has been at this geographic location long enough, so that whatever caused the symptoms had enough time to do it. In other words, we have to have "steady state" conditions.

Consistent with this approach, we now describe the following generic situation. Let us take two unrelated units whose paths overlapped in a particular geographic region [1]. In Figure 1, the grids occupied over a long period of time are shown for two military units; for simplicity, they are denoted here as UIC1 and UIC2, respectively. One easily sees that these two units overlap in two grids, DC and CD. Since by assumptions the causes from both sets of grids contribute the same symptoms and diagnoses, and since both units spend a very long time in their respective grids, the cause and effect are entirely due to these grids for these units. Therefore, to simplify matters, we treat units UIC1 and UIC2 as one compound unit, UIC-Compound, as indicated on Figure 1.

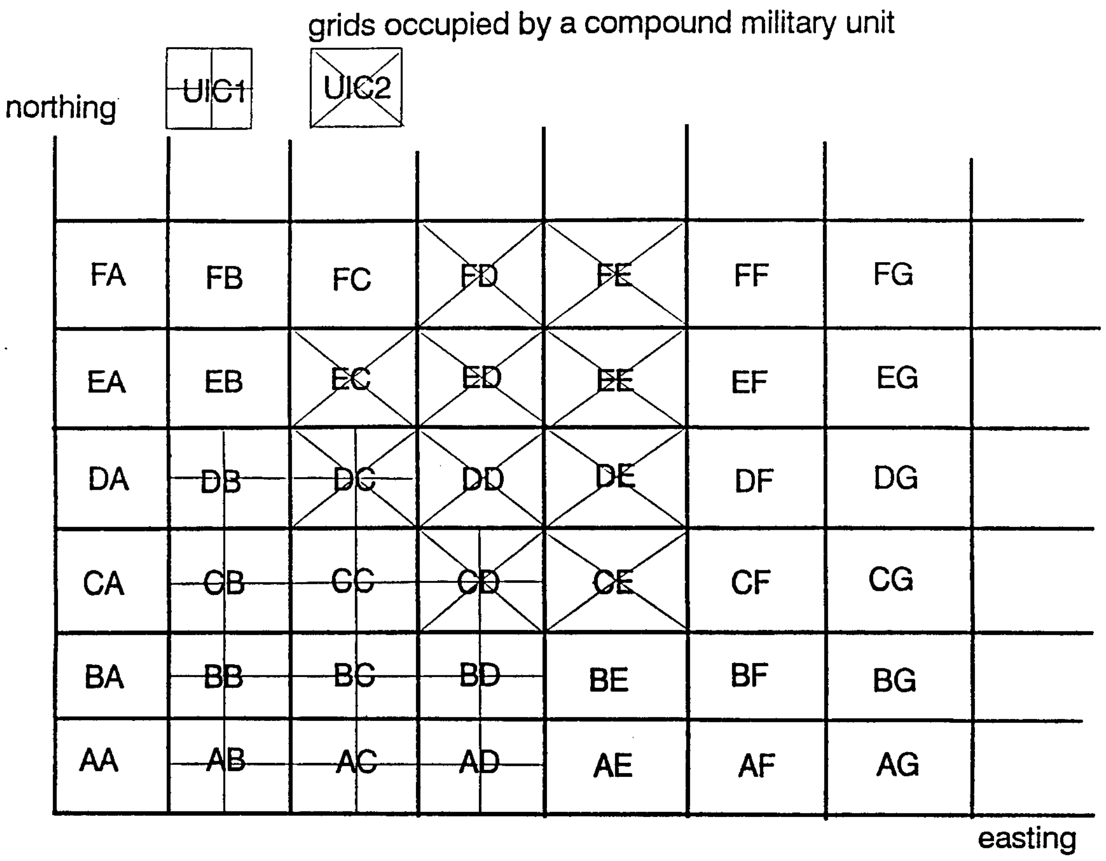

Figure 1. Grids occupied by the compound military unit (UIC-Compound) composed of two separate military units (UIC1 and UIC2).

3. PROBABILISTIC FORMULATION

Once we know what compound unit (UIC-Compound) and geographic region we are dealing with, we can give the probabilistic formulation of the syndrome. At first, we shall only be interested with the number of generalized symptoms, which, as we recall, actually means symptoms plus diagnoses. Specifically, we wish to record the number of people who have, regardless of the type, one generalized symptom, two generalized symptoms, and so on up to six generalized symptoms. For example, if one soldier has just one symptom ICDSYM = 78999 but the other has ICDIAG = 71940, they both qualify as each having just one generalized symptom. On the more formal level, we count every soldier as having one generalized symptom whose six-component generalized symptom/diagnosis vectors look like $\phi = (s_1, 0, 0; 0, 0, 0)$ or $\phi = (0, 0, 0; d_1, 0, 0)$ and denote their number as n_1. Similarly, with two generalized symptoms, we count every soldier whose six-component generalized symptom/diagnosis vectors look like $\phi = (s_1, s_2, 0; 0, 0, 0)$, $\phi = (s_1, 0, 0; d_1, 0, 0)$, $\phi = (0, 0, 0; d_1, d_2, 0)$; we denote the number of soldiers as n_2, and so on all the way up to n_6, which is the number of soldiers whose six-component generalized symptom/diagnosis vectors look like $\phi = (s_1, s_2, s_3; d_1, d_2, d_3)$ with none of the components being equal to zero (of course, different soldiers in principle will have different s_1's, etc.). Finally, we can write this down again in the six-component vector form as

$$n = (n_1, n_2, ..., n_6). \tag{4}$$

Now we remember that we have introduced the continuous generalized symptom variable x, which for our purposes varies between 0 and 6. The reason we want this continuous generalized symptom variable is that, after having found $n_1, n_2, ..., n_6$, we shall be able to construct analytical soldier probability distribution functions in terms of the continuous generalized symptom variable x. Namely, what we are talking about here is the probability that a soldier selected randomly from the UIC-Compound has the generalized symptom number x. However, to connect to a real situation, we have to relate this x to the discrete number.

Now the minimum and maximum numbers of generalized symptoms that a person can have are, respectively, 1 and 6. Consequently, from this continuous x, we define the discrete number of symptoms, say, x_i, as follows:

$$1 \le x \le 1.5 \rightarrow x_1 = 1;\ 1.5 < x \le 2.5 \rightarrow x_2 = 2;\ 2.5 < x \le 3.5 \rightarrow x_3 = 3;$$
$$3.5 < x \le 4.5 \rightarrow x_4 = 4;\ 4.5 < x \le 5.5 \rightarrow x_5 = 5;\ \rightarrow 5.5 < x \le 6 \rightarrow x_6 = 6. \tag{5}$$

We see that actually $x_i = i$; $i = 1, 2, ..., 6$. So, for example, if $x = 2.3$, we say that the number of generalized symptoms a randomly chosen soldier has is two.

Now we are ready to construct the probability distribution function $P(x; n_1, n_2, ..., n_6)$, telling us the probability that a randomly chosen soldier has a number of generalized symptoms x; in a more precise language, what one has here is that $P(x; n_1, n_2, ..., n_6)dx$ is the probability that a randomly chosen soldier has a generalized symptom number between x and $(x + dx)$. We will, however, use the "sloppy" language that $P(x; n_1, n_2, ..., n_6)$ is simply the "probability of getting x" for a randomly chosen soldier. Next comes the important question of normalization of $P(x; n_1, n_2, ..., n_6)$. While the physical domain of x is clearly between 0 and 6, we, unfortunately, have to settle to be just between 1 and 6. The reason for this is that there will be no records of soldiers, who, although feeling the "effects" of the some illness, under examination showed zero generalized symptoms. Hence, our probability distribution function will be defined for x satisfying $1 \le x \le 6$ with the normalization:

$$\int_{1}^{6} P(x;\ n_1,\ n_2,\ ...,\ n_6)dx\ =\ 1\,. \tag{6}$$

Finally, we are ready to construct $P(x;\ n_1, n_2, ..., n_6\ )$. To do that, we use the finite element method, whose rather rigorous exposition can be found in [2], while a much simpler exposition, which we follow here, is in [3].

Here the continuous generalized symptom variable x is a one-dimensional variable. The nodal points, which are needed to obtain the interpolation functions for constructing $P(x;\ n_1, n_2, ..., n_6)$, are then simply where x takes the physical discrete values:

$$x_i = i,\ i = 1, 2, ..., 6. \tag{7}$$

These, in essence, define the one-dimensional equidistant six-nodes fifth power element [2, 3] with which are associated six interpolation functions and which we will derive. For their derivation, we need the six-component polynomial basis vector [2, 3]:

$$b(x) = (1, x, x^2, x^3, x^5), \tag{8}$$

whose form tells us why this element is the fifth power element. With the help of (8), in the "vector of the vector" form, we next write down the 6 × 6 matrix $\underline{m}$

$$\underline{m}\ =\ \begin{pmatrix} b(x_1) \\ b(x_2) \\ b(x_3) \\ b(x_4) \\ b(x_5) \\ b(x_6) \end{pmatrix}, \tag{9}$$

whose detailed expression one can easily obtain from (9), (8), and (7). However, to obtain the interpolation functions, we actually need the inverse of this matrix. With the help of Mathematica [4], for example, we obtain:

$$\underline{m}^{-1} = \begin{pmatrix} 6 & -15 & 20 & -15 & 6 & -1 \\ -87/10 & 117/4 & -127/3 & 33 & -27/2 & 137/60 \\ 29/6 & -461/24 & 31 & -307/12 & 65/6 & -15/8 \\ -31/24 & 137/24 & -121/12 & 107/12 & -95/24 & 17/24 \\ 1/6 & -19/24 & 3/2 & -17/12 & 2/3 & -1/8 \\ -1/120 & 1/24 & -1/12 & 1/12 & -1/24 & 1/120 \end{pmatrix}. \tag{10}$$

According to the finite element method [2, 3], the six interpolation functions, $u_1(x)$, $u_2(x)$, ..., $u_6(x)$, are obtained in the vector form from contracting the polynomial basis vector $b(x)$ with $\underline{m}^{-1}$:

$$u(x) = (u_1(x), u_2(x), \ldots, u_6(x)) = b(x).\underline{m}^{-1}, \tag{11}$$

yielding specifically,

$$u_1(x) = 6 - \frac{87x}{10} + \frac{29x^2}{6} - \frac{31x^3}{24} + \frac{x^4}{6} - \frac{x^5}{120}, \tag{12a}$$

$$u_2(x) = -15 + \frac{117x}{4} - \frac{461x^2}{24} + \frac{137x^3}{24} - \frac{19x^4}{24} + \frac{x^5}{24}, \tag{12b}$$

$$u_3(x) = 20 - \frac{127x}{3} + 31x^2 - \frac{121x^3}{12} + \frac{3x^4}{2} - \frac{x^5}{12}, \tag{12c}$$

$$u_4(x) = -15 + 33x - \frac{307x^2}{12} + \frac{107x^3}{12} - \frac{17x^4}{12} + \frac{x^5}{12}, \tag{12d}$$

$$u_5(x) = 6 - \frac{27x}{2} + \frac{65x^2}{6} - \frac{95x^3}{24} + \frac{2x^4}{3} - \frac{x^5}{24}, \tag{12e}$$

$$u_6(x) = -1 + \frac{137x}{60} - \frac{15x^2}{8} + \frac{17x^3}{24} - \frac{x^4}{8} + \frac{x^5}{120}. \tag{12f}$$

With the help of Mathematica [4] or directly with some hard labor, one can see that

$$u_4(x) = u_3(-x+7),\ u_5(x) = u_2(-x+7),\ u_6(x) = u_1(-x+7); \tag{13}$$

as a consequence, on Figures 2, 3, and 4, we show, respectively, only $u_1(x)$, $u_2(x)$, and $u_3(x)$. One should note that each $u_i(x)$ has the largest (positive) value at its node $x_i = i$ going through zeros at other nodes and changing signs with smaller values in between. As such, they are poised to be continuous approximation for histograms centered at $x_i = i$, whose widths are given as

$$\Delta x_1 = (x_2 - \frac{1}{2}) - x_1;\ \Delta x_i = (x_{i+1} - \frac{1}{2}) - (x_{i-1} + \frac{1}{2}),\ i = 2, 3, ..., 5;\ \Delta x_6 = x_6 - (x_5 + \frac{1}{2}), \tag{14}$$

where relation (5) was taken into account.

Furthermore, the functions $u_i(x)$ also satisfy

$$\sum_{i=1}^{6} u_i(x) = 1. \tag{15}$$

Defining

$$I_i = \int_1^6 u_i(x)\,dx,\ i = 1, 2, ..., 6, \tag{16}$$

we write down the results in the vector form:

$$I = (I_1 = 0.32986, I_2 = 1.30208, I_3 = 0.86806, I_4 = I_3, I_5, = I_2, I_6 = I_1), \tag{17}$$

where we notice that (17) and (13) are consistent with each other. Because of (15), we clearly have

$u_1(x) = u_6(-x+7)$

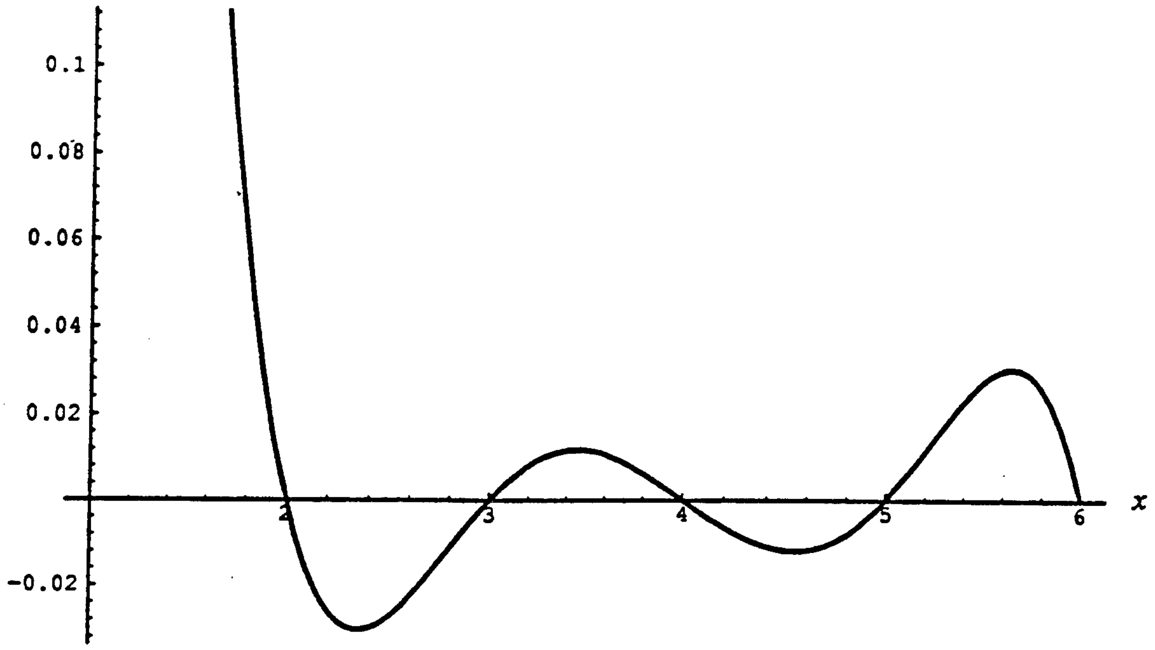

Figure 2. Interpolation function associated with the first (and sixth) nodal point.

$u_2(x) = u_5(-x+7)$

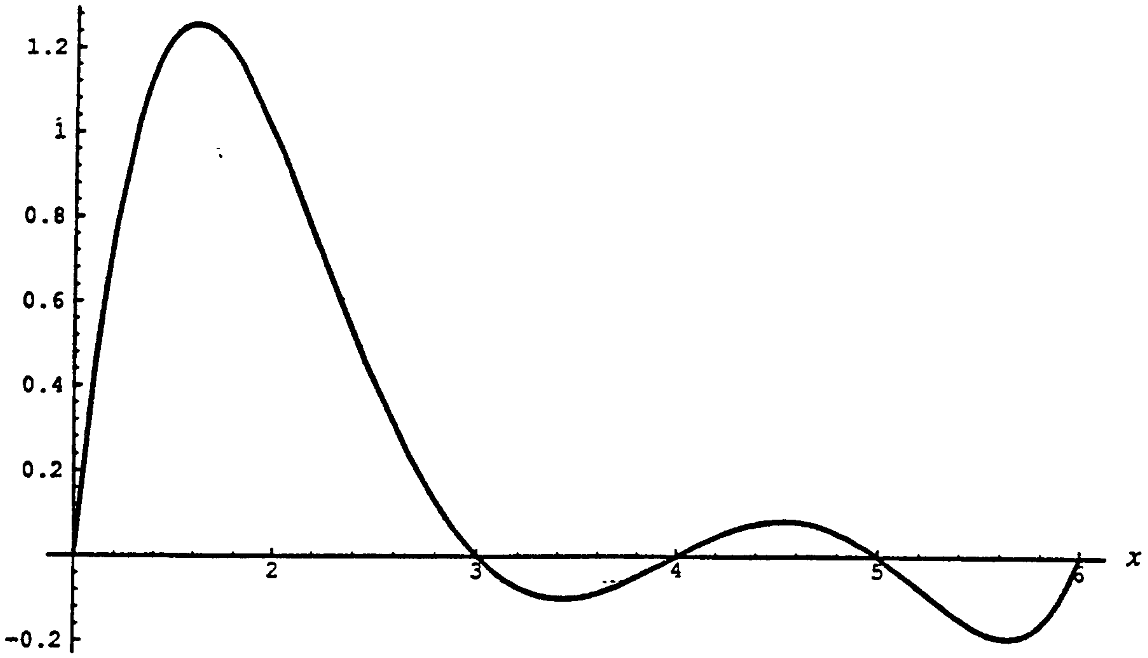

Figure 3. Interpolation function associated with the second (and fifth) nodal point.

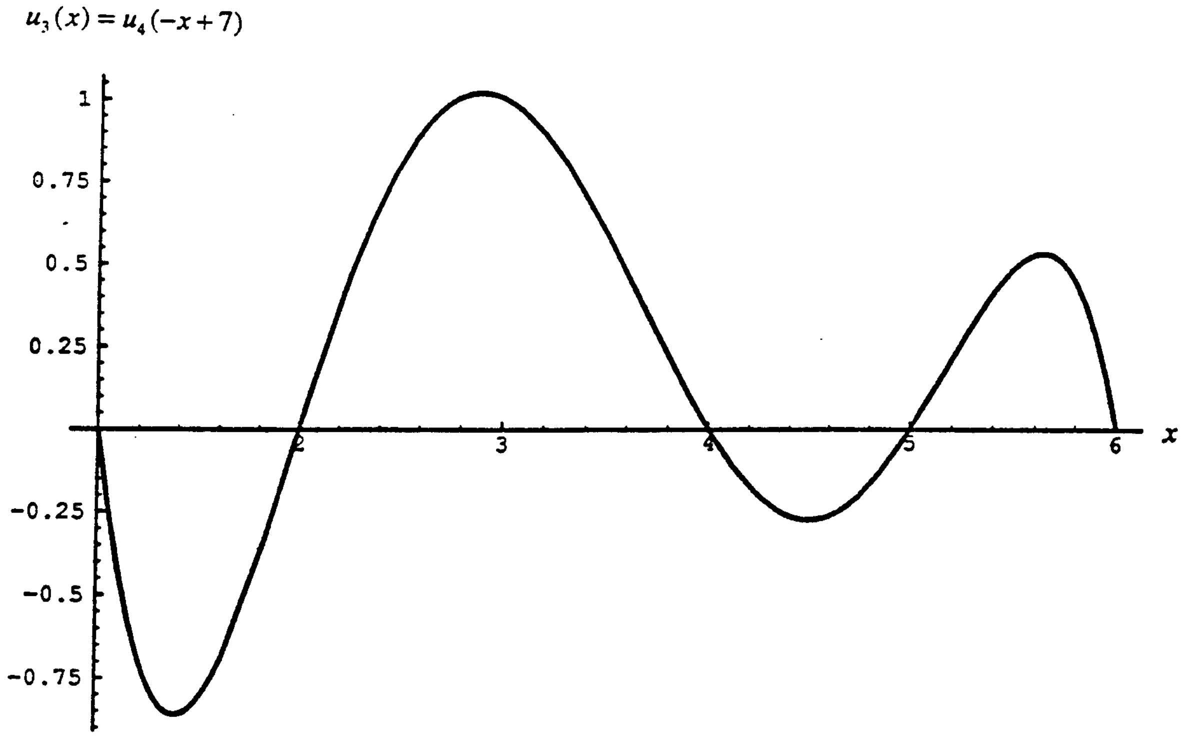

Figure 4. Interpolation function associated with the third (and fourth) nodal point.

$$\sum_{i=1}^{6} I_i = \int_{1}^{6} dx = 5. \tag{18}$$

Now it is easily seen that the normalized probability distribution function $P(x; n_1, n_2, ..., n_6)$ is given as a ratio of $u(x).n$ to $I.n$;

$$P(x); n_1, n_2, ..., n_6 = \frac{n.u(x)}{n.I} = \frac{\sum_{i=1}^{6} n_i u_i(x)}{\sum_{j=1}^{6} n_j I_j}. \tag{19a,b}$$

So the only thing we need to know now are the number of soldiers, n_i, that are associated with the number of generalized symptoms, $i = 1, 2, ..., 6$, and we know the probability for a randomly chosen soldier of having x generalized symptoms.

A word of caution. The number of people, n_i, that are associated with the number of generalized symptoms, $i = 1, 2, ..., 6$, may redefine the range of the continuous generalized symptoms. Namely, suppose that $n_1 = n_2 = 0$ and the other n's are different from zero. Then the range of x rather than being from 1 to 6 will be only over Δx_3, Δx_4, Δx_5, and Δx_6 combined range. The distribution function is obtained from (19b) by setting $n_1 = n_2 = 0$. The new distribution function is now approximately normalized over the new combined range. The reason for this is that, numerically, the I_j are almost independent of the range of integration. For example, if we define

$$I_i' = \int_{\Delta x_i} u_i(x)\,dx, \; i = 1, 2, ..., 6,$$

where Δx_i are defined by (14), by direct numerical comparisons, one deduces that $I_i' \cong I_i$. Hence, it appears quite plausible that enlarging the range of integration will not make much difference. At this time we do not, however, discuss the probability functions with "holes," i.e., when some n_i in the middle of the whole range are equal to zero.

However, establishing the probability distribution function is not the end of the story. A quantity that potentially has very good use is the generalized symptom density $xsy(x;\, n_1, n_2, ..., n_6)$ and is defined as

$$xsy(x;\, n_1, n_2, ..., n_6) = xP(x;\, n_1, n_2, ..., n_6). \tag{20a}$$

Now, directly from (19b), we then have,

$$xsy(x;\, n_1,\, n_2,\, ...,\, n_6 = \frac{\sum_{i=1}^{6} n_i\, x\, u_i(x)}{\sum_{j=1}^{6} n_j I_j}. \tag{20b}$$

We will use this density to divide people into very ill and less ill later on. However, the average number of generalized symptoms over all randomly chosen people is simply the integral over (17b):

$$<xsy(x;\ n_1,\ n_2,\ ...,\ n_6)> \ = \int_1^6 dx\ xsy(x;\ n_1,\ n_2,\ ...,\ n_6 \ = \frac{\sum_{i=1}^{6} n_i J_i}{\sum_{j=1}^{6} n_j I_j}, \qquad (21)$$

where

$$J_i = \int_1^6 dx\, x\, u_i(x). \qquad (22)$$

As we see, two of the quantities that we have to know are J_i and n_i, which are easily calculable. Next, in the six-component vector form, we list the J_i from (22):

$$J = (J_1 = 0.46627, J_2 = 1.92212, J_3 = 3.96825, J_4 = 2.10814, J_5 = 7.19246, J_6 = 1.84276). \qquad (23)$$

Again, combining (23) and (15), we obtain the sum rule

$$\sum_{i=1}^{6} J_i = \int_1^6 x\, dx = \frac{35}{2}. \qquad (24)$$

One easily verifies the correctness of relations (19) and (21) by assuming that $n_1 = n_2 = ... = n_6 \equiv n$. We have that $P(x;\ n,\ n,\ ...,\ n) = 1/5$ and $< xsy(x;\ n,\ n,\ ...,) > = 7/2$, i.e., results indicate that any x will occur with equal probability, as it should.

From our discussion after relation (19) we remember that, for example, if $n_1 = n_2 = 0$ and the other n's are different from zero, then the range of x rather than being from 1 to 6 will be only over Δx_3, Δx_4, Δx_5, and Δx_6 combined range. There we argued that since approximately $I_i' \cong I_i$, the distribution function is obtained from (19b) by setting $n_1 = n_2 = 0$. However, the situation here is a little bit different since $< xsy(x;\ 0, 0, n_3, n_4, ..., n_6) >$ involves also the J_i. Defining again

$$J_i' = \int_{\Delta x_i} x\ \ u_i(x)\, dx,\ i = 1, 2, ..., 6,$$

we again see that numerically, although not perfect, we still can write $J_i \cong J_i'$. Consequently, even in the new smaller range, we can use the expression (21) when evaluating $< xsy(x;\ 0, 0, n_3, n_4, ..., n_6) >$.

We essentially have what is needed to do some numerical exercises.

Application Examples. The idea here is very simple. From the derived probability distribution functions, we compare statistical properties of different UIC-Compounds occupying different geographic regions. If these statistical properties are very much the same for different UIC-Compounds at their respective geographic areas, then the geographies and what is on them have very little to do with the syndrome. On the other hand, if these statistical properties are very different at different geographic sites, the geographies and what is on them play an important role for each of these UIC-Compound's syndromes. We should remember, however, that the assumption of sufficiently long time exposures to the geographies, we made at the beginning, should hold.

So we assume a particular UIC-Compound associated with a particular geography. From construction of the generalized symptom vector for each ill soldier, as described in section 2, we obtain the components of the vector n. Here, as an example, we assume to be of the form:

$$n = (80, 50, 40, 30, 20, 5), \tag{25}$$

that is, $n_1 = 80$ is the number of people whose generalized symptom/diagnosis vector looks like $\phi = (s_1,\ 0,\ 0;\ 0,\ 0,\ 0)$ or $\phi = (0,\ q_1,\ 0;\ d\ ,\ 0,\ 0)$; etc. Now, the probability distribution function $P(x;\ 80, 50, 40, 30, 20, 5)$ can be written explicitly from relations (19b) and (17). Its form is shown on Figure 5. We see that the choice of the six-component n-vector is rather reasonable, since from relation (21) we have that the average number of generalized symptoms is numerically $< xsy(x;\ 80, 50, 40, 30, 20, 5) > =$

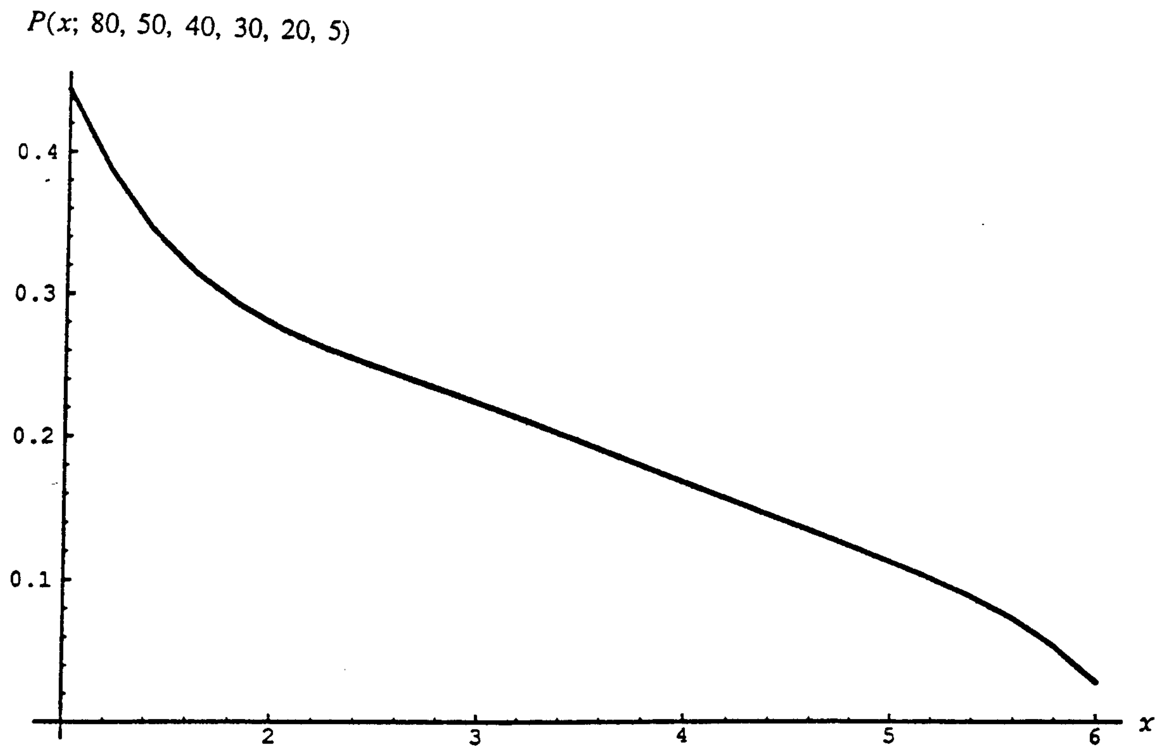

Figure 5. Example plot of the probability distribution function vs. the continuous generalized symptom variable x.

2.826, which, according to our assignments, means that a randomly chosen person would have about three generalized symptoms.

Relation (20) yields the generalized symptom density *xsy(x*; 80, 50, 40, 30, 20, 5) itself. Figure 6 shows its form. The curve has a maximum at about $x_m \cong 3.4$. The significance of x_m is that it divides people into two categories, i.e., those with $x < x_m$, which, by definition, are just ill, and those with $x > x_m$, which, by definition, are very ill. These definitions allow us to deduce the syndrome itself by restricting ourselves to just very ill people. Hence, in this example, we pull from the database only those people who have 4, 5, and 6 generalized symptoms. Indiscriminately, these generalized symptoms are identified according to the ICD-9-CM codes, as described in section 2.1. The program should be written which ranks them according to the frequency of their occurrence, however, now separately for ordinary symptoms and separately for diagnoses. Combining the most frequent symptoms and diagnoses, say, three from each (these numbers are not fixed; we could easily take four from each, etc.), we can deduce the syndrome either from the NIH sets of syndromes, as described in section 2.1, or simply by giving our most frequent symptoms and diagnoses to

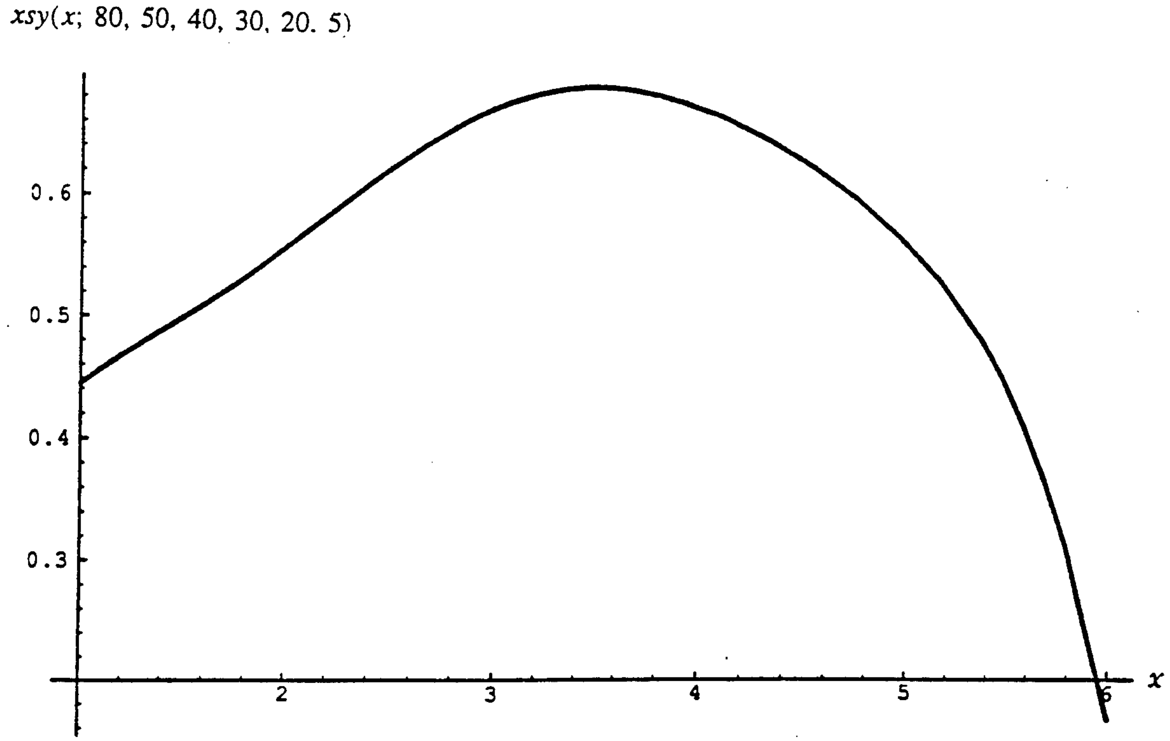

Figure 6. Example plot of the generalized symptom density vs. the continuous generalized symptom variable *x*.

a team of competent and independent physicians who should decide on the causes (syndrome), or both. The result should be a syndrome associated with our particular probability distribution function.

Now going from one geographic area with a UIC-Compound to another geographic area with a different UIC-Compound, we should be able to make a direct comparison between the probability distribution functions, the generalized symptom densities, the average numbers of generalized symptoms, and the x_m's. If these sets of quantities do not differ very much for the two geographies, then the syndrome at both places is the same and geographies have very little to do with it. If, however, they do differ significantly for the two geographies, then the geographies themselves should hold the key as to why people at one place are sick in one way and at the other place the other way.

4. DISCUSSION AND CONCLUSION

It is too late to try to apply the probabilistic methods described here to the Gulf War Syndrome question. However, in the future, we believe this method should be able to answer rather straightforwardly whether the constructed probability distribution functions for every UIC or UIC-Compound look alike or not, i.e., whether the geography and the details of the battlefield environment have something to do with the syndrome or not. Of course, if the probability distribution functions are all very much alike, then the cause of the syndrome is something that the UICs and UIC-Compounds did to themselves. Finally, if desired, for each probability distribution function, one can derive the syndrome with the help of generalized symptom density and the database, as described in the previous section.

5. REFERENCES

1. Gluckman, A. G. "A Statistical Study Correlating the Reported Cases of Gulf Syndrome to Battlefield Locations of Afflicted U.S. Army Personnel During the Iraq-Kuwait War." ARL-TR-800, U.S. Army Research Laboratory, Adelphi, MD, July 1995.

2. Dhatt, G., and G. Touzot. The Finite Element Method Displayed. NY: Wiley, 1984.

3. Šoln, J. Z. "Finite Element, Finite Difference, and Finite Volume Methods: Examples and Their Comparisons." ARL-TR-979, U.S. Army Research Laboratory, Adelphi, MD, March 1996.

4. Wolfram, S. Mathematica. Reading, MA: Addison-Wesley, 2nd Ed., 1991.

INTENTIONALLY LEFT BLANK.

NO. OF COPIES	ORGANIZATION
2	DEFENSE TECHNICAL INFO CTR ATTN DTIC DDA 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	HQDA DAMO FDQ ATTN DENNIS SCHMIDT 400 ARMY PENTAGON WASHINGTON DC 20310-0460
1	US MILITARY ACADEMY MATH SCI CTR OF EXCELLENCE DEPT OF MATHEMATICAL SCI ATTN MDN A MAJ DON ENGEN THAYER HALL WEST POINT NY 10996-1786
1	DIRECTOR US ARMY RESEARCH LAB ATTN AMSRL CS AL TP 2800 POWDER MILL RD ADELPHI MD 20783-1145
1	DIRECTOR US ARMY RESEARCH LAB ATTN AMSRL CS AL TA 2800 POWDER MILL RD ADELPHI MD 20783-1145
3	DIRECTOR US ARMY RESEARCH LAB ATTN AMSRL CI LL 2800 POWDER MILL RD ADELPHI MD 20783-1145
	ABERDEEN PROVING GROUND
2	DIR USARL ATTN AMSRL CI LP (305)

NO. OF
COPIES ORGANIZATION

1 OSD OUSD AT
STRT TAC SYS
ATTN DR SCHNEITER
3090 DEFNS PENTAGON RM 3E130
WASHINGTON DC 20301-3090

1 ASST SECY ARMY RESEARCH
DEVELOPMENT ACQUISITION
ATTN SARD ZD RM 2E673
103 ARMY PENTAGON
WASHINGTON DC 20310-0103

1 ASST SECY ARMY RESEARCH
DEVELOPMENT ACQUISITION
ATTN SARD ZP RM 2E661
103 ARMY PENTAGON
WASHINGTON DC 20310-0103

1 ASST SECY ARMY RESEARCH
DEVELOPMENT ACQUISITION
ATTN SARD ZS RM 3E448
103 ARMY PENTAGON
WASHINGTON DC 20310-0103

1 ASST SECY ARMY RESEARCH
DEVELOPMENT ACQUISITION
ATTN SARD ZT RM 3E374
103 ARMY PENTAGON
WASHINGTON DC 20310-0103

1 UNDER SEC OF THE ARMY
ATTN DUSA OR
RM 2E660
102 ARMY PENTAGON
WASHINGTON DC 20310-0102

1 ASST DEP CHIEF OF STAFF
OPERATIONS AND PLANS
ATTN DAMO FDZ RM 3A522
460 ARMY PENTAGON
WASHINGTON DC 20310-0460

1 DEPUTY CHIEF OF STAFF
OPERATIONS AND PLANS
ATTN DAMO SW RM 3C630
400 ARMY PENTAGON
WASHINGTON DC 20310-0400

NO. OF
COPIES ORGANIZATION

1 ARMY RESEARCH LABORATORY
ATTN AMSRL SL
PROGRAMS AND PLANS MGR
WSMR NM 88002-5513

1 ARMY RESEARCH LABORATORY
ATTN AMSRL SL E
MR MARES
WSMR NM 88002-5513

1 ARMY TRADOC ANL CTR
ATTN ATRC W
MR KEINTZ
WSMR NM 88002-5502

1 ARMY TRNG & DOCTRINE CMND
ATTN ATCD B
FT MONROE VA 23651

ABERDEEN PROVING GROUND

1 CDR USATECOM
ATTN: AMSTE-TA

2 DIR USAMSAA
ATTN: AMXSY-ST
AMXSY-D

4 DIR USARL
ATTN: AMSRL-SL,
J WADE (433)
M STARKS (433)
AMSRL-SL-C, J BEILFUSS (E3331)
AMSRL-SL-B, P DEITZ (328)

1 CDR CBDCOM
ATTN: TECHNICAL LIBRARY
BLDG E3330

1 DIR CBIAC
BLDG E3330, RM 150

NO. OF
COPIES ORGANIZATION

1 HQDA
DUSA OR
ROOM 2E660
102 ARMY PENTAGON
WASHINGTON DC 20310-0102

1 DEP CHIEF OF STAFF
OPERATIONS AND PLANS
ATTN DAMO SW ROOM 3C630
400 ARMY PENTAGON
WASHINGTON DC 20310-0400

1 COMMANDER
US ARMY MISSILE COMMAND
ATTN AMSTA CG
REDSTONE ARSENAL AL 35898-5000

1 DIRECTOR
US ARMY BALLISTIC MIS DEFNS
SYS CMND
ATTN ADVNCD TECHLGY CTR
PO BOX 1500
HUNTSVILLE AL 35807-3801

2 COMMANDER
USA STRTGC DEFNS CMND
ATTN CSSD SL C
CSSD SL S
HUNTSVILLE AL 35807-3801

1 COMMANDER
CECOM R&D TECH LIB
ATTN ASQNC ELC IS L R
FORT MONMOUTH NJ 07703-5000

1 COMMANDER
US ARMY RESEARCH OFFICE
ATTN TECH LIB
PO BOX 12211
RESEARCH TRIANGLE PARK NC 27709-2211

1 COMMANDER
USA LOGISTICS MGMT CTR
ATTN DEFNS LGSTCS STUDIES
FORT LEE VA 23801

1 COMMANDER
US ARMY NGIC
ATTN AMXST MC 3
220 SEVENTH ST NE
CHARLOTTESVILLE VA 22901-5396

NO. OF
COPIES ORGANIZATION

1 DIRECTOR
USARL
ATTN AMSRL OP TL
WATERTOWN MA 02172-0001

7 DIRECTOR
USARL
AMSRL SL EA R SHELBURNE
AMSRL SL EG J PALOMO
AMSRL SL EM R FLORES
AMSRL SL EP D ALVAREZ
AMSRL SL ES T ATHERTON
AMSRL SL EV DR K MORRISON
AMSRL SL CA R SUTHERLAND
WSMR NM 88002-5513

1 DIRECTOR
USARL
ATTN AMSRL EP
M V GELNOVATCH
FT MONMOUTH NJ 07703-5601

1 DIRECTOR
USARL
ATTN AMSRL EP EF
DR M DUTTA
FT MONMOUTH NJ 07703-5601

1 DIRECTOR
USARL
ATTN AMSRL SS V DEMONTE
ADELPHI MD 20783-1197

1 DIRECTOR
USARL
ATTN AMSRL SS IC
DR P EMMERMAN
ADELPHI MD 20783-1197

1 DIRECTOR
USARL
ATTN AMSRL SS IC
R WINKLER
ADELPHI MD 20783-1197

1 DIRECTOR
USARL
ATTN AMSRL WT NH
DR H E BRANDT
ADELPHI MD 20783-1197

NO. OF COPIES	ORGANIZATION
1	COMMANDANT
	USA CMND AND GENL STAFF
	FORT LEAVENWORTH KS 66027
1	DIRECTOR
	INSTITUTE OF DEFNS ANLYS
	ATTN LIBRARY
	1801 BEAUREGARD ST
	ALEXANDRIA VA 22311
	ABERDEEN PROVING GROUND
1	DIR, USAERDEC
	ATTN SCBRD RT
1	CDR, USACBDA
	ATTN AMSCB CII
1	CDR, USAATC
	ATTN STECS
76	DIR, USARL
	ATTN AMSRL SL J SMITH
	AMSRL SL BA
	J MORRISSEY
	J HANES
	R KUNKEL
	L ROACH
	AMSRL SL BG A YOUNG
	AMSRL SL BS
	D BELY
	T KLOPCIC
	AMSRL SL BV
	R SANDMEYER
	W MERMAGEN JR
	AMSRL SL C
	J SEIGH
	MAJ GILMAN
	AMSRL SL CM
	D FARENWALD
	B RUTH
	L DAVIS
	L DELICIO
	R JOLLIFFE
	D SLOOP
	R TYTUS
	R ZUM BRUNNEN
	E FIORAVANTE
	M MAR
	DR ŠOLN (30 CPS)
	AMSRL SL CO
	D BAYLOR
	R PROCHAZKA
	AMSRL SL CS
	J BEILFUSS
	T FLORY
	B SMITH
	M SMITH
	J FRANZ
	T MAK
	J NEALON
	R PARSONS
	M KAUFMAN
	D MANYAK
	R POLIMADEI
	M BUMBAUGH
	J CAPOBIANCO
	D DAVIS
	AMSRL SL I
	R REITZ
	D BASSETT
	A YOUNG
	M VOGEL
	E PANUSKA
	D HASKELL
	J FEENEY
	R ZIGLER

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project(0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE December 1996	3. REPORT TYPE AND DATES COVERED Final, June 1995–June 1996

4. TITLE AND SUBTITLE Probabilistic Modeling of a Syndrome	5. FUNDING NUMBERS 14626187480
6. AUTHOR(S) Josip Z. Šoln	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: AMSRL-SL-CM Aberdeen Proving Ground, MD 21005-5068	8. PERFORMING ORGANIZATION REPORT NUMBER ARL–TR–1268
9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)	10.SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.	12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We propose a probabilistic methodology for deducing a syndrome or syndromes (possibly induced by chemical/biological agents) associated with a large number of people from certain geographic areas that have well-established diagnoses and symptoms. Here, using the finite element method and the databases of symptoms and diagnoses, for each geographic area an analytical probability distribution function is established, which gives a probability that a person has a certain number of symptoms/diagnoses. This, in turn, allows us to write down an analytic expression for the symptoms/diagnoses density from which, with the help of databases, one deduces the overall most numerous symptoms and diagnoses; as such, they define the syndrome for the particular geographic area. Now, comparing the syndromes to each other, one can see to what extent geography, and what is on it, affects the syndromes associated with different geographic areas.

14. SUBJECT TERMS modeling, probability, syndrome	15. NUMBER OF PAGES 29
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL

INTENTIONALLY LEFT BLANK.

USER EVALUATION SHEET/CHANGE OF ADDRESS

This Laboratory undertakes a continuing effort to improve the quality of the reports it publishes. Your comments/answers to the items/questions below will aid us in our efforts.

1. ARL Report Number/Author ARL-TR-1268 (Soln) Date of Report December 1996

2. Date Report Received

3. Does this report satisfy a need? (Comment on purpose, related project, or other area of interest for which the report will be used.)

4. Specifically, how is the report being used? (Information source, design data, procedure, source of ideas, etc.)

5. Has the information in this report led to any quantitative savings as far as man-hours or dollars saved, operating costs avoided, or efficiencies achieved, etc? If so, please elaborate.

6. General Comments. What do you think should be changed to improve future reports? (Indicate changes to organization, technical content, format, etc.)

CURRENT ADDRESS

Organization

Name E-mail Name

Street or P.O. Box No.

City, State, Zip Code

7. If indicating a Change of Address or Address Correction, please provide the Current or Correct address above and the Old or Incorrect address below.

OLD ADDRESS

Organization

Name

Street or P.O. Box No.

City, State, Zip Code

(Remove this sheet, fold as indicated, tape closed, and mail.)

(DO NOT STAPLE)

DEPARTMENT OF THE ARMY

OFFICIAL BUSINESS

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO 0001,APG,MD

POSTAGE WILL BE PAID BY ADDRESSEE

DIRECTOR
US ARMY RESEARCH LABORATORY
ATTN AMSRL SL CM
ABERDEEN PROVING GROUND MD 21010-5423